# Identification of polymorphic tandem repeats by direct comparison of genome sequence from different bacterial strains: a web-based resource.

France Denoeud, Gilles Vergnaud

BMC Bioinformatics

BioMed Central

Database



# Identification of polymorphic tandem repeats by direct comparison of genome sequence from different bacterial strains : a web-based resource

France Denœud*[1] and Gilles Vergnaud[1,2]

Address: [1]Laboratoire GPMS, Institut de Génétique et Microbiologie, Bat 400, Université Paris-Sud, 91405 Orsay cedex, France and [2]Centre d'Etudes du Bouchet, BP3, 91710 Vert le Petit, France

Email: France Denœud* - France.Denoeud@igmors.u-psud.fr; Gilles Vergnaud - Gilles.Vergnaud@igmors.u-psud.fr

* Corresponding author

Published: 12 January 2004

*BMC Bioinformatics* 2004, **5**:4

Received: 24 September 2003
Accepted: 12 January 2004

This article is available from: http://www.biomedcentral.com/1471-2105/5/4



## Abstract

**Background:** Polymorphic tandem repeat typing is a new generic technology which has been proved to be very efficient for bacterial pathogens such as *B. anthracis*, *M. tuberculosis*, *P. aeruginosa*, *L. pneumophila*, *Y. pestis*. The previously developed tandem repeats database takes advantage of the release of genome sequence data for a growing number of bacteria to facilitate the identification of tandem repeats. The development of an assay then requires the evaluation of tandem repeat polymorphism on well-selected sets of isolates. In the case of major human pathogens, such as *S. aureus*, more than one strain is being sequenced, so that tandem repeats most likely to be polymorphic can now be selected *in silico* based on genome sequence comparison.

**Results:** In addition to the previously described general Tandem Repeats Database, we have developed a tool to automatically identify tandem repeats of a different length in the genome sequence of two (or more) closely related bacterial strains. Genome comparisons are pre-computed. The results of the comparisons are parsed in a database, which can be conveniently queried over the internet according to criteria of practical value, including repeat unit length, predicted size difference, etc. Comparisons are available for 16 bacterial species, and the orthopox viruses, including the variola virus and three of its close neighbors.

**Conclusions:** We are presenting an internet-based resource to help develop and perform tandem repeats based bacterial strain typing. The tools accessible at http://minisatellites.u-psud.fr now comprise four parts. The Tandem Repeats Database enables the identification of tandem repeats across entire genomes. The Strain Comparison Page identifies tandem repeats differing between different genome sequences from the same species. The "Blast in the Tandem Repeats Database" facilitates the search for a known tandem repeat and the prediction of amplification product sizes. The "Bacterial Genotyping Page" is a service for strain identification at the subspecies level.

## Background

Molecular epidemiology, the integration of molecular typing and conventional epidemiological studies, is likely to add significant value to analyses of infections caused by pathogenic bacteria (see [1] for review). Multilocus Sequence Typing (MLST) for instance is now a major reference method for the molecular epidemiology of *Neisseria meningitidis* and other human pathogens [2]. In this

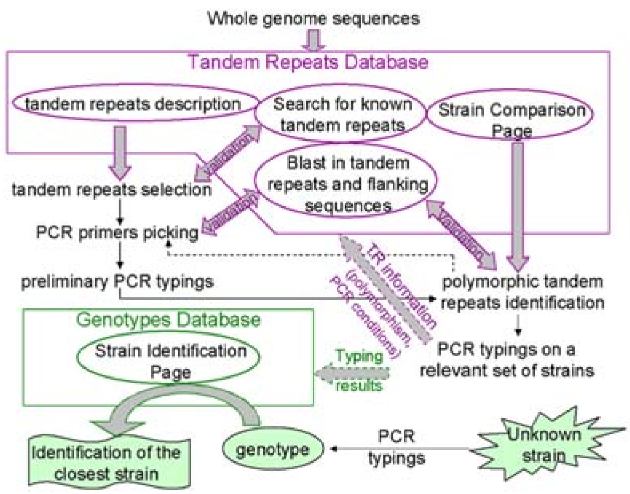
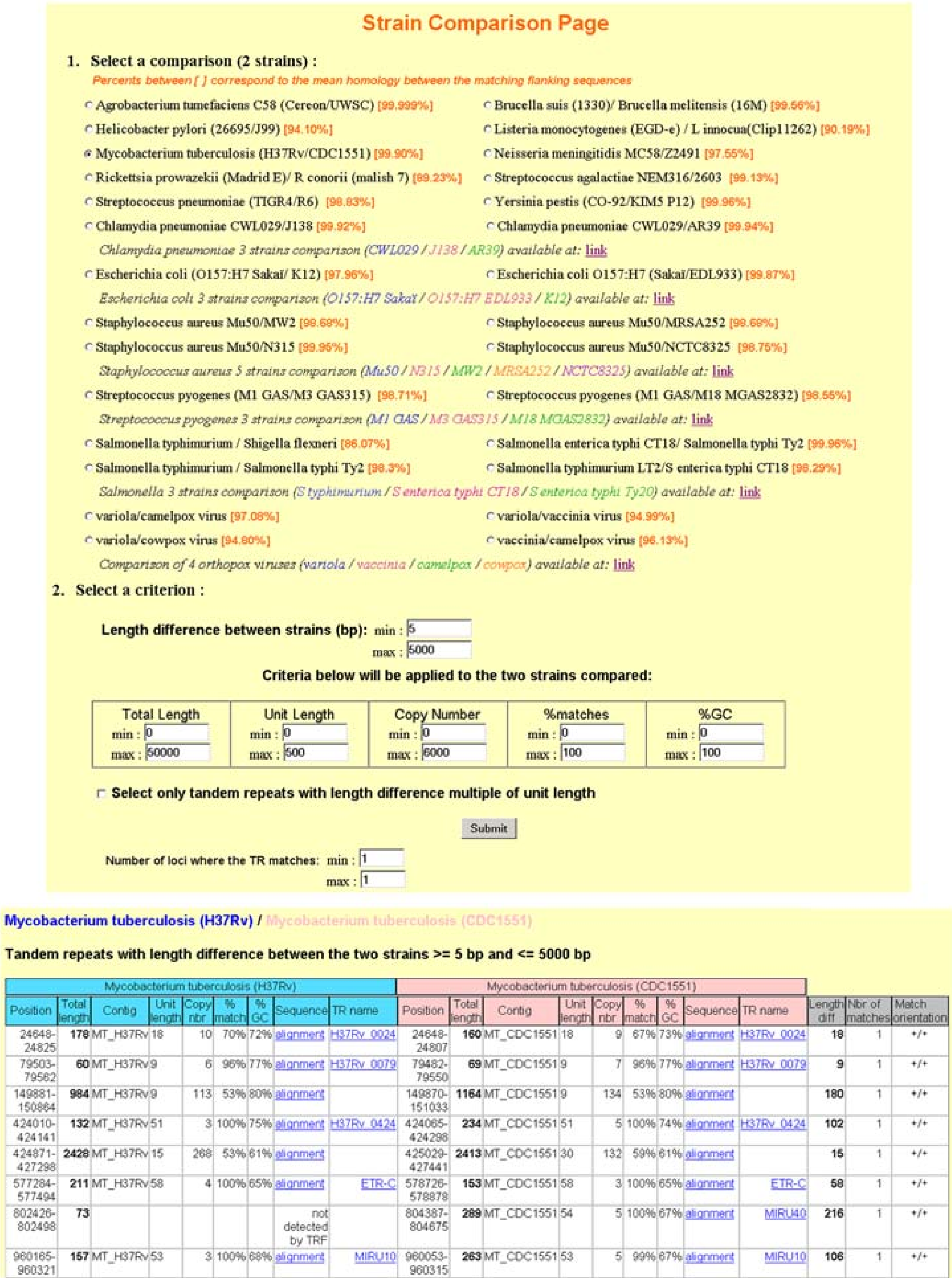



**Figure 1**
The procedure to find polymorphic tandem repeats for use in strain typing. The steps leading from the release of a complete (or incomplete) genome sequence to the validation of new polymorphic markers are described. The purpose of the web-based tools developed is to facilitate the bioinformatics data-management steps.

kind of assay, a set of typically 7 genes is partially sequenced, and the resulting data is converted into sequence types, which can be easily stored in databases, and compared to others. However a number of significant pathogens, including *M. tuberculosis* [3], *B. anthracis* and *Y. pestis* [4] are not amenable to this approach, because of the recent emergence of these pathogens and the resulting rarity of sequence variations. In these pathogens, tandem repeats (TRs) are a source of very informative markers for strain genotyping [5-10]. Tandem repeats in pathogenic bacteria were initially identified within genes associated with bacterial virulence [11,12]. In other instances, the contribution of tandem repeats to genome polymorphism was established after extensive searches based for instance on AFLP (amplified fragment length polymorphism) profiling. This is well illustrated by *B. anthracis*, in which polymorphic bands in AFLP patterns [13] were subsequently demonstrated by sequencing to be due to tandem repeat variations [14]. Eventually, some of these tandem repeats have been shown to directly contribute to phenotypic variations of the *B. anthracis* exosporium which makes the outer layer of the spores [15]. The frequent observation that tandem repeat-containing genes are often associated with outer membrane proteins suggests that such genes help bacteria adapt to their environment, and may be to some extent mutation hotspots as a result of positive selection.

| Comparison | %matches between flanking sequences | % of flanking seqs not rearranged | % of polymorphic tandem repeats | positions of matching tandem repeats |
|---|---|---|---|---|
| *Agrobacterium tumefaciens* C58 Cereon/UWSC | 99.99 | 99.95 | 0 | |
| *Salmonella enterica typhi* CT18/Ty2 | 99.96 | 98.83 | 2.04 | |
| *Yersinia pestis* CO-92/KIM5 P12 | 99.96 | 98.51 | 8.47 | |
| *Staphylococcus aureus* Mu50/N315 | 99.95 | 98.05 | 1.38 | |
| *Chlamydia pneumoniae* CWL029/AR39 | 99.94 | 99.86 | 1.41 | |
| *Chlamydia pneumoniae* CWL029/J138 | 99.92 | 99.86 | 2.30 | |
| *Mycobacterium tuberculosis* H37Rv/CDC1551 | 99.90 | 99.50 | 4.06 | |
| *Escherichia coli* O157:H7 Sakai/EDL933 | 99.87 | 99.49 | 2.13 | |
| *Brucella suis* 1330/ *melitensis* 16M | 99.56 | 99.08 | 7.85 | |
| *Streptococcus agalactiae* NEM316/2603 | 99.13 | 85.44 | 2.76 | |
| *Streptococcus pneumoniae* TIGR4/R6 | 99.83 | 90.01 | 16.50 | |
| *Staphylococcus aureus* Mu50/NCTC8325 | 98.75 | 92.35 | 8.12 | |
| *Streptococcus pyogenes* M1GAS/M3GAS | 98.71 | 86.30 | 6.38 | |
| *Staphylococcus aureus Mu50/MW2* | 98.68 | 91.04 | 8.24 | |
| *Staphylococcus aureus* Mu50/MRSA252 | 98.68 | 90.17 | 8.81 | |
| *Streptococcus pyogenes* M1GAS/M18GAS | 98.55 | 89.19 | 5.82 | |
| *Salmonella typhimurium* LT2 / typhi CT18 | 98.29 | 87.65 | 5.44 | |
| *Escherichia coli* O157:H7 Sakai/ K12 | 97.96 | 77.23 | 8.92 | |
| *Neisseria meningitidis* MC58/Z2491 | 97.54 | 92.66 | 18.88 | |
| *Helicobacter pylori* 26695/J99 | 94.10 | 87.23 | 18.91 | |
| *Listeria monocytogenes* EGDe / *innocua* Clip | 90.19 | 74.13 | 3.99 | |
| *Rickettsia prowazekii* Madrid E / *conorii* malish7 | 89.23 | 61.05 | 8.23 | |
| *Salmonella typhimurium* / *Shigella flexneri* | 86.06 | 47.16 | 6.23 | |

**Figure 2**
**Comparison of strains using different indexes.** The four columns correspond to (from left to right): (1) mean %identity provided by BLAST when the match occurred on more than half the length of the 500 bp of submitted flanking sequence ; (2) proportion (%) of flanking sequences that matched on more than half their length between the two strains ; (3) proportion (%) of tandem repeats of a different size in the two strains ; and (4) plot of the positions of homologous tandem repeat loci in the two genomes which indirectly reflects large scale genome rearrangements. Species are listed according to the first index (mean %identity)



## Strain Comparison Page

**1. Select a comparison (2 strains) :**

*Percents between [ ] correspond to the mean homology between the matching flanking sequences*

- Agrobacterium tumefaciens C58 (Cereon/UWSC) [99.999%]
- Brucella suis (1330)/ Brucella melitensis (16M) [99.56%]
- Helicobacter pylori (26695/J99) [94.10%]
- Listeria monocytogenes (EGD-e) / L innocua(Clip11262) [90.19%]
- Mycobacterium tuberculosis (H37Rv/CDC1551) [99.90%]
- Neisseria meningitidis MC58/Z2491 [97.55%]
- Rickettsia prowazekii (Madrid E)/ R conorii (malish 7) [89.23%]
- Streptococcus agalactiae NEM316/2603 [99.13%]
- Streptococcus pneumoniae (TIGR4/R6) [99.83%]
- Yersinia pestis (CO-92/KIM5 P12) [99.96%]
- Chlamydia pneumoniae CWL029/J138 [99.92%]
- Chlamydia pneumoniae CWL029/AR39 [99.94%]

*Chlamydia pneumoniae 3 strains comparison (CWL029 / J138 / AR39) available at: link*

- Escherichia coli (O157:H7 Sakai/ K12) [97.96%]
- Escherichia coli O157:H7 (Sakai/EDL933) [99.87%]

*Escherichia coli 3 strains comparison (O157:H7 Sakai / O157:H7 EDL933 / K12) available at: link*

- Staphylococcus aureus Mu50/MW2 [99.69%]
- Staphylococcus aureus Mu50/MRSA252 [99.69%]
- Staphylococcus aureus Mu50/N315 [99.95%]
- Staphylococcus aureus Mu50/NCTC8325 [98.75%]

*Staphylococcus aureus 5 strains comparison (Mu50 / N315 / MW2 / MRSA252 / NCTC8325) available at: link*

- Streptococcus pyogenes (M1 GAS/M3 GAS315) [98.71%]
- Streptococcus pyogenes (M1 GAS/M18 MGAS2832) [98.55%]

*Streptococcus pyogenes 3 strains comparison (M1 GAS / M3 GAS315 / M18 MGAS2832) available at: link*

- Salmonella typhimurium / Shigella flexneri [86.07%]
- Salmonella enterica typhi CT18/ Salmonella typhi Ty2 [99.96%]
- Salmonella typhimurium / Salmonella typhi Ty2 [98.3%]
- Salmonella typhimurium LT2/S enterica typhi CT18 [98.29%]

*Salmonella 3 strains comparison (S typhimurium / S enterica typhi CT18 / S enterica typhi Ty20) available at: link*

- variola/camelpox virus [97.08%]
- variola/vaccinia virus [94.99%]
- variola/cowpox virus [94.80%]
- vaccinia/camelpox virus [96.13%]

*Comparison of 4 orthopox viruses (variola / vaccinia / camelpox / cowpox) available at: link*

**2. Select a criterion :**

**Length difference between strains (bp):** min : 5 max : 5000

**Criteria below will be applied to the two strains compared:**

| Total Length | Unit Length | Copy Number | %matches | %GC |
|---|---|---|---|---|
| min : 0 max : 50000 | min : 0 max : 500 | min : 0 max : 6000 | min : 0 max : 100 | min : 0 max : 100 |

**Select only tandem repeats with length difference multiple of unit length**

Submit

**Number of loci where the TR matches:** min : 1 max : 1

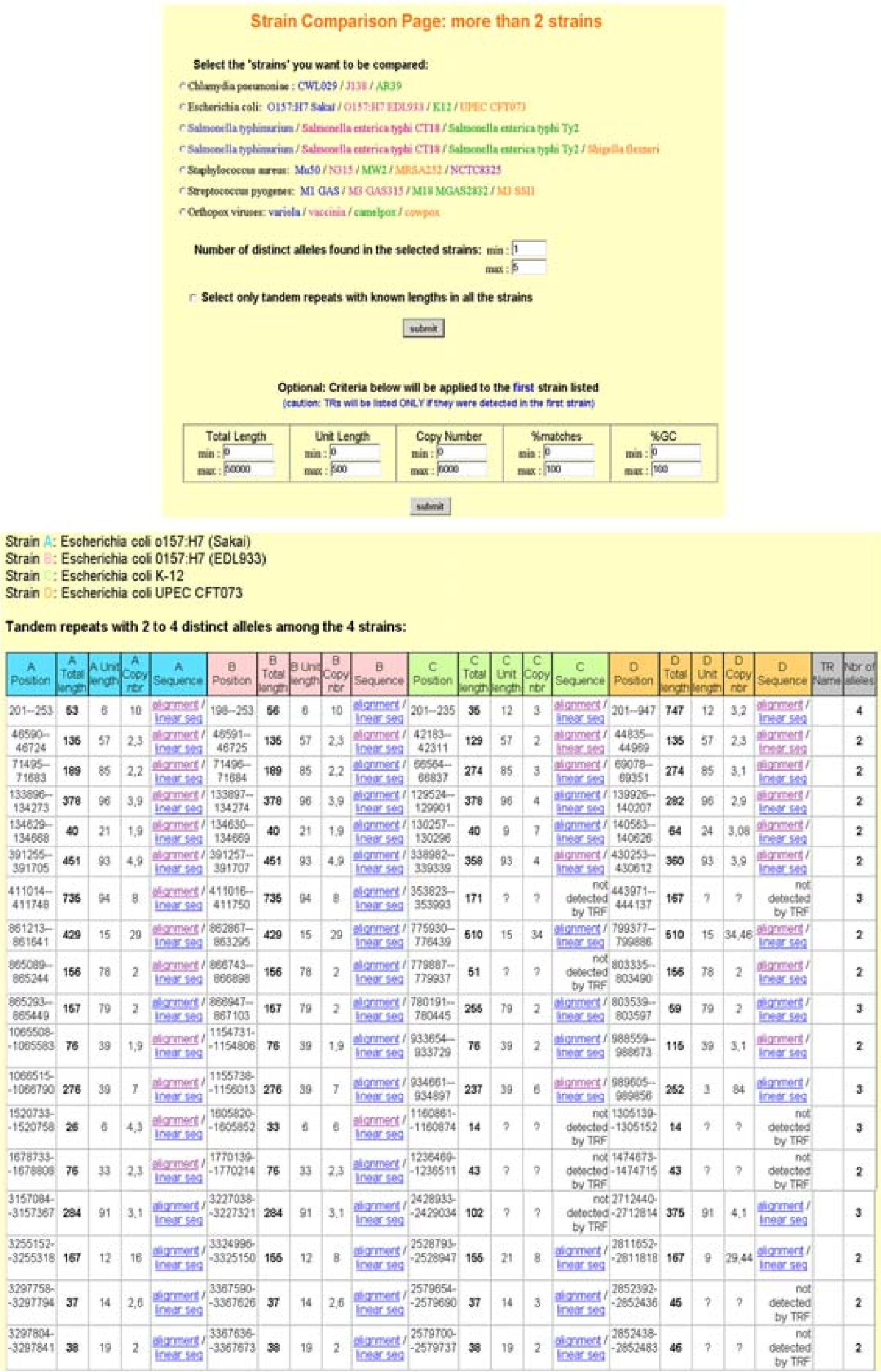
**Mycobacterium tuberculosis (H37Rv) / Mycobacterium tuberculosis (CDC1551)**

**Tandem repeats with length difference between the two strains >= 5 bp and <= 5000 bp**

| Mycobacterium tuberculosis (H37Rv) | | | | | | | | | Mycobacterium tuberculosis (CDC1551) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Position | Total length | Contig | Unit length | Copy nbr | % match | % GC | Sequence | TR name | Position | Total length | Contig | Unit length | Copy nbr | % match | % GC | Sequence | TR name | Length diff | Nbr of matches | Match orientation |
| 24648-24825 | **178** | MT_H37Rv | 18 | 10 | 70% | 72% | alignment | H37Rv_0024 | 24648-24807 | **160** | MT_CDC1551 | 18 | 9 | 67% | 73% | alignment | H37Rv_0024 | **18** | 1 | +/+ |
| 79503-79562 | **60** | MT_H37Rv | 9 | 6 | 96% | 77% | alignment | H37Rv_0079 | 79482-79550 | **69** | MT_CDC1551 | 9 | 7 | 96% | 77% | alignment | H37Rv_0079 | **9** | 1 | +/+ |
| 149881-150864 | **984** | MT_H37Rv | 9 | 113 | 53% | 80% | alignment | | 149870-151033 | **1164** | MT_CDC1551 | 9 | 134 | 53% | 80% | alignment | | **180** | 1 | +/+ |
| 424010-424141 | **132** | MT_H37Rv | 51 | 3 | 100% | 75% | alignment | H37Rv_0424 | 424065-424298 | **234** | MT_CDC1551 | 51 | 5 | 100% | 74% | alignment | H37Rv_0424 | **102** | 1 | +/+ |
| 424871-427298 | **2428** | MT_H37Rv | 15 | 268 | 53% | 61% | alignment | | 425029-427441 | **2413** | MT_CDC1551 | 30 | 132 | 59% | 61% | alignment | | **15** | 1 | +/+ |
| 577284-577494 | **211** | MT_H37Rv | 58 | 4 | 100% | 65% | alignment | ETR-C | 578726-578878 | **153** | MT_CDC1551 | 58 | 3 | 100% | 65% | alignment | ETR-C | **58** | 1 | +/+ |
| 802426-802498 | **73** | | | | | | not detected by TRF | | 804387-804675 | **289** | MT_CDC1551 | 54 | 5 | 100% | 67% | alignment | MIRU40 | **216** | 1 | +/+ |
| 960165-960321 | **157** | MT_H37Rv | 53 | 3 | 100% | 68% | alignment | MIRU10 | 960053-960315 | **263** | MT_CDC1551 | 53 | 5 | 99% | 67% | alignment | MIRU10 | **106** | 1 | +/+ |

**Figure 3**

Example of a query in the Strain Comparison Page. On the top, the query page shows the 28 comparisons currently available (others will be added as new genome sequences are finished and released). Bottom, the result of a query performed for *Mycobacterium tuberculosis* strains H37Rv and CDC1551 is summarized.



Strain Comparison Page: more than 2 strains

Select the 'strains' you want to be compared:

- Chlamydia pneumoniae : CWL029 / J138 / AR39
- Escherichia coli: O157:H7 Sakai / O157:H7 EDL933 / K12 / UPEC CFT073
- Salmonella typhimurium / Salmonella enterica typhi CT18 / Salmonella enterica typhi Ty2
- Salmonella typhimurium / Salmonella enterica typhi CT18 / Salmonella enterica typhi Ty2 / Shigella flexneri
- Staphylococcus aureus: Mu50 / N315 / MW2 / MRSA252 / NCTC8325
- Streptococcus pyogenes: M1 GAS / M3 GAS315 / M18 MGAS8232 / M3 SSI1
- Orthopox viruses: variola / vaccinia / camelpox / cowpox

Number of distinct alleles found in the selected strains: min : 1 max : 5

Select only tandem repeats with known lengths in all the strains

submit

Optional: Criteria below will be applied to the first strain listed
(caution: TRs will be listed ONLY if they were detected in the first strain)

| Total Length | Unit Length | Copy Number | %matches | %GC |
|---|---|---|---|---|
| min : 0 max : 50000 | min : 0 max : 500 | min : 0 max : 5000 | min : 0 max : 100 | min : 0 max : 100 |

submit

Strain A: Escherichia coli o157:H7 (Sakai)
Strain B: Escherichia coli 0157:H7 (EDL933)
Strain C: Escherichia coli K-12
Strain D: Escherichia coli UPEC CFT073

Tandem repeats with 2 to 4 distinct alleles among the 4 strains:

| A Position | A Total length | A Unit length | A Copy nbr | A Sequence | B Position | B Total length | B Unit length | B Copy nbr | B Sequence | C Position | C Total length | C Unit length | C Copy nbr | C Sequence | D Position | D Total length | D Unit length | D Copy nbr | D Sequence | TR Name | Nbr of alleles |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 201--253 | **53** | 6 | 10 | alignment / linear seq | 198--253 | **56** | 6 | 10 | alignment / linear seq | 201--235 | **35** | 12 | 3 | alignment / linear seq | 201--947 | **747** | 12 | 3,2 | alignment / linear seq | | **4** |
| 46590--46724 | **136** | 57 | 2,3 | alignment / linear seq | 46591--46725 | **135** | 57 | 2,3 | alignment / linear seq | 42183--42311 | **129** | 57 | 2 | alignment / linear seq | 44835--44969 | **135** | 57 | 2,3 | alignment / linear seq | | **2** |
| 71495--71683 | **189** | 85 | 2,2 | alignment / linear seq | 71496--71684 | **189** | 85 | 2,2 | alignment / linear seq | 66564--66837 | **274** | 85 | 3 | alignment / linear seq | 69078--69351 | **274** | 85 | 3,1 | alignment / linear seq | | **2** |
| 133896--134273 | **378** | 96 | 3,9 | alignment / linear seq | 133897--134274 | **378** | 96 | 3,9 | alignment / linear seq | 129524--129901 | **378** | 96 | 4 | alignment / linear seq | 139926--140207 | **282** | 96 | 2,9 | alignment / linear seq | | **2** |
| 134629--134668 | **40** | 21 | 1,9 | alignment / linear seq | 134630--134669 | **40** | 21 | 1,9 | alignment / linear seq | 130257--130296 | **40** | 9 | 7 | alignment / linear seq | 140563--140626 | **64** | 24 | 3,08 | alignment / linear seq | | **2** |
| 391255--391705 | **451** | 93 | 4,9 | alignment / linear seq | 391257--391707 | **451** | 93 | 4,9 | alignment / linear seq | 338982--339339 | **358** | 93 | 4 | alignment / linear seq | 430253--430612 | **360** | 93 | 3,9 | alignment / linear seq | | **2** |
| 411014--411748 | **735** | 94 | 8 | alignment / linear seq | 411016--411750 | **735** | 94 | 8 | alignment / linear seq | 353823--353993 | **171** | ? | ? | not detected by TRF | 443971--444137 | **167** | ? | ? | not detected by TRF | | **3** |
| 861213--861641 | **429** | 15 | 29 | alignment / linear seq | 862867--863295 | **429** | 15 | 29 | alignment / linear seq | 775930--776439 | **510** | 15 | 34 | alignment / linear seq | 799377--799886 | **510** | 15 | 34,46 | alignment / linear seq | | **2** |
| 865089--865244 | **156** | 78 | 2 | alignment / linear seq | 866743--866898 | **156** | 78 | 2 | alignment / linear seq | 779887--779937 | **51** | ? | ? | not detected by TRF | 803335--803490 | **156** | 78 | 2 | alignment / linear seq | | **2** |
| 865293--865449 | **157** | 79 | 2 | alignment / linear seq | 866947--867103 | **157** | 79 | 2 | alignment / linear seq | 780191--780445 | **255** | 79 | 2 | alignment / linear seq | 803539--803597 | **59** | 79 | 2 | alignment / linear seq | | **3** |
| 1065508--1065583 | **76** | 39 | 1,9 | alignment / linear seq | 1154731--1154806 | **76** | 39 | 1,9 | alignment / linear seq | 933654--933729 | **76** | 39 | 2 | alignment / linear seq | 988559--988673 | **115** | 39 | 3,1 | alignment / linear seq | | **2** |
| 1066515--1066790 | **276** | 39 | 7 | alignment / linear seq | 1155738--1156013 | **276** | 39 | 7 | alignment / linear seq | 934661--934897 | **237** | 39 | 6 | alignment / linear seq | 989605--989856 | **252** | 3 | 84 | alignment / linear seq | | **3** |
| 1520733--1520758 | **26** | 6 | 4,3 | alignment / linear seq | 1605820--1605852 | **33** | 6 | 6 | alignment / linear seq | 1160861--1160874 | **14** | ? | ? | not detected by TRF | 1305139--1305152 | **14** | ? | ? | not detected by TRF | | **3** |
| 1678733--1678808 | **76** | 33 | 2,3 | alignment / linear seq | 1770139--1770214 | **76** | 33 | 2,3 | alignment / linear seq | 1236469--1236511 | **43** | ? | ? | not detected by TRF | 1474673--1474715 | **43** | ? | ? | not detected by TRF | | **2** |
| 3157084--3157367 | **284** | 91 | 3,1 | alignment / linear seq | 3227038--3227321 | **284** | 91 | 3,1 | alignment / linear seq | 2428933--2429034 | **102** | ? | ? | not detected by TRF | 2712440--2712814 | **375** | 91 | 4,1 | alignment / linear seq | | **3** |
| 3255152--3255318 | **167** | 12 | 16 | alignment / linear seq | 3324996--3325150 | **155** | 12 | 8 | alignment / linear seq | 2528793--2528947 | **155** | 21 | 8 | alignment / linear seq | 2811652--2811818 | **167** | 9 | 29,44 | alignment / linear seq | | **2** |
| 3297758--3297794 | **37** | 14 | 2,6 | alignment / linear seq | 3367590--3367626 | **37** | 14 | 2,6 | alignment / linear seq | 2579654--2579690 | **37** | 14 | 3 | alignment / linear seq | 2852392--2852436 | **45** | ? | ? | not detected by TRF | | **2** |
| 3297804--3297841 | **38** | 19 | 2 | alignment / linear seq | 3367636--3367673 | **38** | 19 | 2 | alignment / linear seq | 2579700--2579737 | **38** | 19 | 2 | alignment / linear seq | 2852438--2852483 | **46** | ? | ? | not detected by TRF | | **2** |

**Figure 4**
Example of a query in the Strain Comparison Page for more than two strains. Top, the query page shows the 6 comparisons currently available (others will be added as new genome sequences are finished and released). Bottom, the result of a query performed for *Escherichia coli* strains O157:H7 Sakaï, O157:H7 EDL933, K12 and UPEC-CFT073 is summarized. In several loci, the size of the repeat is listed differently for the different strains, which is due to different detections by the Tandem Repeats Finder, usually as a result of internal variations within the tandem array. Total length is calculated from positions of matching flanking sequences in the different strains, and does not necessarily correspond to the length of the tandem repeat detected by TRF in the locus. "Number of alleles" refers to the number of predicted sizes differing by at least 5 bp among the strains compared.

Polymorphic tandem repeats (VNTRs, for Variable Number of Tandem Repeats), once identified, provide convenient tools requiring ordinary molecular biology equipment and the data can be easily exchanged and compared. The resulting assay, called MLVA (for multiple locus VNTR analysis) can even be automated [16]. We have developed tools which facilitate the bioinformatics step of genome analysis required to start a project. A previously described Tandem Repeats Database enables the identification of tandem repeats across entire genomes [9,10,17-19]. It has been constantly updated, with now more than a hundred bacterial genomes available, compared to 35 at the onset of the database. We present here a new and major development of this resource which takes advantage of the fact that more than two different strains from the same species have now been sequenced at least for a number of major human pathogens. As a result, the tools accessible over the Internet at http://minisatellites.u-psud.fr now comprise four complementary parts. The newly added resource, the Strain Comparison Page, takes advantage of the availability of genome sequences from more than one strain from a growing number of species to directly identify tandem repeats differing between the sequenced strains. This is of interest because the vast majority of tandem repeats is often not polymorphic [19]. The "Blast in the Tandem Repeats Database" page facilitates the search for a known tandem repeat, the prediction of PCR amplification products size, and the verification of primer specificity. Once an MLVA assay has been set up, and carefully validated by typing collections of isolates, it is relatively easy to construct databases of genotypes to be used locally or which can be queried across the Internet. The "Bacterial Genotyping Page" illustrates a freely accessible, fast and easy to use internet-based service for strain comparisons, in which a user can compare a genotype produced for one of his isolates to the existing data.

## Construction and content

### *The Tandem Repeats Database main page*

Tandem repeats were identified from finished microbial genome sequences (as listed by the Genome OnLine Database [20]) using the tandem repeats finder (TRF) software [21,22] with the following options: alignment parameters, "2,3,5" (these parameters are the less stringent ones), minimum alignment score to report repeat, 50 (this score allows to detect short structures), maximum period size, 500 base-pairs. When the program reported redundant (overlapping) repeats, the redundancy was eliminated as described in [23], before import in the database. The database uses Microsoft Access 2000 and the querying process uses Active Server Pages (ASP, Microsoft) with Perlscripts or VBscripts. Perl was obtained from the ActiveState Programmer Network [24]. The database is hosted on a server running under Windows 2000 server (Microsoft). The tandem repeats database main page is described in more detail in [9].

### *The Strain Comparison page*

Sequence comparisons used BLAST [25]. The BLAST software was obtained from the NCBI FTP site [26]. The flanking sequences of TRs from one strain were compared to the whole sequence of the other strain (and reciprocally, to avoid missing some tandem repeats that would not appear in the tandem repeats database for one strain because they were not detected by the Tandem Repeats Finder [21] -for instance because there is only one copy of the repeated unit in the considered strain). The resulting list of matching tandem repeats was then imported in the database, where it can be queried. The comparison of more than two strains was made possible through a supplemental step before import in the database: the synthesis of several 2-strains comparisons, of the same "reference" strain against each of the others (matching between TRs of the different strains was deduced from the positions on the reference strain).

### *The Blast page*

The Blast Page allows users to run BLAST [25] in the tandem repeats and flanking sequences from the database via Perlscripts. The Blast outputs are linked to the database, in order to easily obtain the description of identified tandem repeats.

### *The Bacterial Genotyping page*

The web-page site performing identifications was developed using the BNserver application (version 3.0, Applied-Maths, Belgium) and ASP (Microsoft) using Perlscript. The typing results (gel images and resulting data) were managed using the Bionumerics software package as described in [10]. The output of a query is a list of strains and genotypes from the database together with similarity scores.

## Utility

### *The procedure to find polymorphic tandem repeats (TRs) for use in strain typing*

Figure 1 shows the steps leading from a genome sequence to the exploitation of polymorphic tandem repeats for bacterial strain genotyping. Although Tandem Repeats are easily identified using the Tandem Repeats Database, TR polymorphism must be evaluated by typing across a set of relevant strains. If the sequences of several strains of the species of interest are available, the Strain Comparison Page can be used to directly identify tandem repeats predicted to be polymorphic in size between the two (or more) sequenced strains. However, it is important to keep in mind that the tandem repeats predicted as being polymorphic will depend on the sequenced strains and well-planned surveys of isolates will still be necessary. The available tools do not replace this validation step, as the value of each marker must be carefully established on an appropriate set of isolates. The definition of an appropriate set of isolates depends upon the question which is being addressed, *i.e.* large scale or local epidemiology. The Blast Page has been implemented in the tandem repeats database in order to easily determine the size of the expected PCR amplification products. The database is also manually updated to contain PCR conditions as well as polymorphism index, and links to the original reports [27] (input from users is welcome). Eventually, when an MLVA assay has been fully developed and validated, typing data can be made accessible so that individual queries can be run. The Bacterial Genotyping Page illustrates how this could work. The genotyping data for a strain can be entered and submitted via this page. The output is the description of the closest strains. The data which has been submitted is not incorporated in the database itself, since this would require stringent data validation steps. In the following sections, we are presenting the web-based resources associated with this procedure.

### The "Strain Comparison" pages

The strain comparison pages are available via [28]. The comparison of two strains is based on a pre-computed BLAST [25] analysis of the flanking sequences of tandem repeats from one strain against the other, and vice-versa. Figure 2 summarizes the results of this first step for 23 comparisons. Three indexes are scored (see figure legend): (1) the "mean %identity" between the flanking sequences is a measure of single nucleotide polymorphism (SNPs) frequency (not insertions-deletions), (2) the proportion (%) of flanking sequences that matched the flanking sequence of its homologue in the other strain on more than half of the 500 bp assayed here – *i.e.* that were not rearranged, by insertion of mobile elements for instance -, (3) the proportion (%) of tandem repeats that were found to be of a different length between the two strains being compared. In addition, the positions of matching tandem repeats in the two genomes is plotted to reveal large-scale genome rearrangements. A number of situations are observed: for instance *Yersinia pestis* orientalis strain CO-92 [29], and medievalis strain KIM5 P12 [30] show a very high "mean %identity" (99.96 %), in agreement with the recent emergence of *Yersinia pestis* [4]. In spite of this, the two strains differ by a high number of large rearrangements (as seen on the plot), which reflects the high genome plasticity observed in this species [31], together with a relatively high rate of polymorphic tandem repeats (8.47%). In contrast, *Listeria monocytogenes* strain EGD-e and *Listeria innocua* strain Clip 11262 have a lower homology (90.19%) and only 3.99% of polymorphic tandem repeats in spite of the evolutionary distance (see Figure 2).

The strain comparison page allows queries in the tandem repeats database according to the tandem repeat length difference between the two strains compared, and also to other tandem repeats characteristics (unit length, copy number, etc...). Figure 3 illustrates a query done for *Mycobacterium tuberculosis* strains H37Rv and CDC1551 [32]: the query "length difference ≥ 5 bp" identifies 58 tandem repeats (8 are shown on Figure 3). This prediction has been tested for the 30 loci amenable to PCR analysis and polymorphism has been confirmed in all cases [10].

When more than two strains have been sequenced, a synthesis of the results of several 2-strains comparisons is also available. Figure 4 illustrates a query made for *Escherichia coli* strains O157:H7 Sakaï, O157:H7 EDL933, K12, and UPEC-CFT073 [33-35]: 87 tandem repeats were found with 2 to 4 alleles among the 4 strains (18 of which are listed in Figure 4).

### The "Blast in the Tandem Repeats Database" page

To facilitate the identification of already studied tandem repeats, we implemented BLAST [25] against the tandem repeats from the database, *i.e.* the tandem repeats themselves and their flanking sequences. The Blast page is available at [36]. All bacteria can be queried at once, which allows the identification of tandem repeats families, conserved in several bacterial species. Another page is dedicated to the Blast of PCR primers and provides the size of the PCR products in all the species/strains where the primers match. Figure 5 shows the results of searching the PCR primer pair from tandem repeat H37Rv_0024_18 bp [10] in all bacteria: as expected, the PCR primer pair matches *Mycobacterium tuberculosis* strains H37Rv and CDC1551, providing different PCR product lengths.

### The Bacterial Genotyping page

The Bacterial Genotyping page [37] provides one illustration on how tandem repeat typing data can be made available via internet to allow external users to query genotyping data (*Bacillus anthracis*, *Yersinia pestis*, *Mycobacterium tuberculosis*, *Pseudomonas aeruginosa* for the moment) and compare a new strain to existing data as previously described in [10]. For each locus, allele sizes can be selected among a list of possibilities (observed sizes). The results of the query indicate a similarity score and include links to the complete data recorded for each strain listed. This page is just meant as an illustration and prototype. MLVA reference data could also be made available for downloading as tabular data files, or can be copied from published datasets, which can then be complemented by in-house data, and analyzed by the appropriate clustering software.

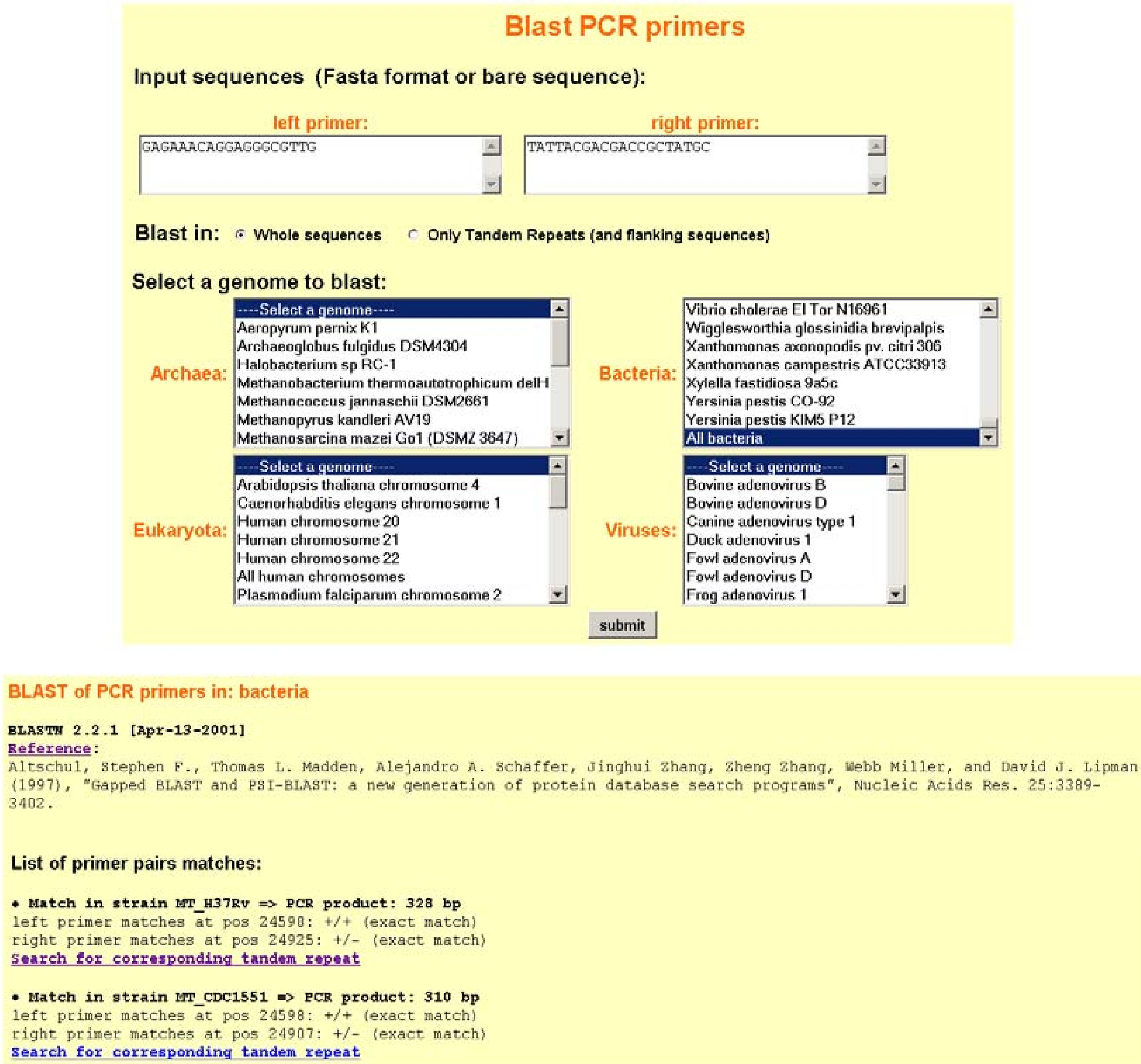



**Figure 5**
Example of a query in the "Blast of PCR primers" page, providing the length of the PCR products in the strains/species where the primer pair matches, and links to the corresponding tandem repeats descriptions.

## Discussion

### *Bacterial genomes evolution*

As shown by the indexes from Figure 2, there are different ways to represent the divergence/similarity between two strains. They are not correlated, suggesting independent evolution processes. First, the "mean %identity" between two genomes reflects point mutations, and is an indicator of the time passed since the two strains diverged. For instance, *Yersinia pestis* is known to be of recent emergence [4] and shows a high "mean %identity" between strains CO-92 (orientalis) and KIM5 P12 (medievalis). In contrast, and as shown by the dot plot, large genome rearrangements occurred in this genome, which is representative of a high genome plasticity [31]. The index "% of flanking sequences not rearranged" is an indicator of small-scale genome rearrangements, such as the insertions of mobile elements. This index is low for genomes rich in mobile elements, like *Streptococcus agalactiae*, in which such elements significantly contribute to strain diversity [38]. Finally, the index "% of polymorphic tandem repeats" between two strains represents the tandem repeats evolution rate. For the moment, the mechanisms of bacterial VNTRs mutations have not been precisely investigated, but it seems likely to be independent of the other processes mentioned, as there are no correlations between the indexes. Figure 2 provides clues to assess which typing method(s) will be efficient in the different species. For instance, the two bacterial species *Salmonella typhimurium* strain LT2 [39] and *Shigella flexneri* strain 2a301 [40] share only 86.06% of sequence identity, clearly making the identification of matching tandem repeats between the two species difficult and of low significance. MLVA analysis appears to be of highest interest for the subspecies typing of highly monomorphic species including *Yersinia pestis*, *Bacillus anthracis*, *Mycobacterium tuberculosis* and *Brucella* [9,10,41].

### *Strain comparison efficiency*

The sequencing of more than one strain for some bacterial species allows direct identification of polymorphic tandem repeats, assuming that no sequencing errors occurred. Earlier investigations provide good reasons to believe that tandem repeats in the size range considered here (a few hundred base-pairs) are correctly sequenced, and consequently, that the strain comparison data is reliable. As a negative control, the comparison of two independent sequences from the same strain of *Agrobacterium tumefaciens* strain (C58), one from Cereon genomics [42] and the other from Washington University [43], shows that no length polymorphism is detected among tandem repeats (Figure 2) between the two independent sequences. As a positive control, the tandem repeats predicted to be polymorphic by genome sequence comparison between the two strains of *M. tuberculosis* have indeed been proved polymorphic by PCR typing of isolates [10].

Selection based on comparison of sequence data from two strains will miss some polymorphic loci. Indeed, the results provided by the approach rely upon the phylogenetic distance between the two strains being compared. If the strains are very closely related, only a few TRs will be found different between them, but these tandem repeats will probably be the most polymorphic ones. Conversely, if the strains are distant in the phylogenetic tree, a larger number of polymorphic TRs will be found, some of them will be only moderately polymorphic. Obviously, when a few well-selected strains have been sequenced, it is likely that very few polymorphic tandem repeats are undetected in the Strain Comparison pages.

It is of course still going to be very important to determine the TR allele frequency for isolates carefully selected to be representative of the global diversity of a given pathogen before suggesting the configuration of an MLVA assay to use in subsequent studies. In addition, those TR markers that are highly polymorphic in diverse test panels of isolates may be monomorphic when applied to isolates responsible for local outbreaks. The configuration of TR markers used to make up an assay needs to be determined empirically with representative local isolates and tailored to the study population and study questions.

### *Polymorphic tandem repeats selection for species with only one sequenced strain*

The identification of simple criteria able to predict tandem repeat polymorphism when genome sequence data is available for only one strain would indeed greatly facilitate the development of MLVA assays. It would seem reasonable for instance to expect that the number of copies and the internal homogeneity of tandem arrays are strong predictors [23]. We take advantage here of the many strain comparisons which are made available via the strain comparison pages to evaluate such criteria.

We have analyzed bacteria with at least three sequenced genomes (*Staphylococcus aureus*: 6 strains, *Escherichia coli*: 4 strains, *Streptococcus pyogenes*: 4 strains and *Salmonella typhi* and *typhimurium*: 3 strains). We assume that in such cases, only a few polymorphic tandem repeats are missed in the comparisons. We compared the distribution of tandem repeats sequence characteristics among the group of "polymorphic" loci (differing in at least two of the strains compared, excluding length differences between strains that resulted from microdeletions in the flanking sequences) and the others. Comparisons were performed for the following sequence characteristics: unit length, copy number, total length, %GC, GC bias (=|%G-%C|/(%G+%C)), %matches, and HistoryR (a score derived from tandem repeat history reconstruction algorithm [44] as described in [23]). None of the variables were normally distributed, as tested with Kolmogorov-Smirnov test, so a non-parametric Wilcoxon test was used to compare the distributions, which were judged significantly different at the .05 level of the statistic (2 tailed). Distributions were significantly different for all 4 species studied for %matches, total length and copy number. As shown on Figure 6, polymorphic TRs have a higher internal conservation and total length than monomorphic ones. Copy number, which is correlated with total length, is also higher among polymorphic TRs.

Selecting the longest and most conserved tandem repeats should thus improve polymorphic TRs identification. Table 1 illustrates the query "total length ≥ 80 bp and %matches ≥ 80%" applied to the four species used to find

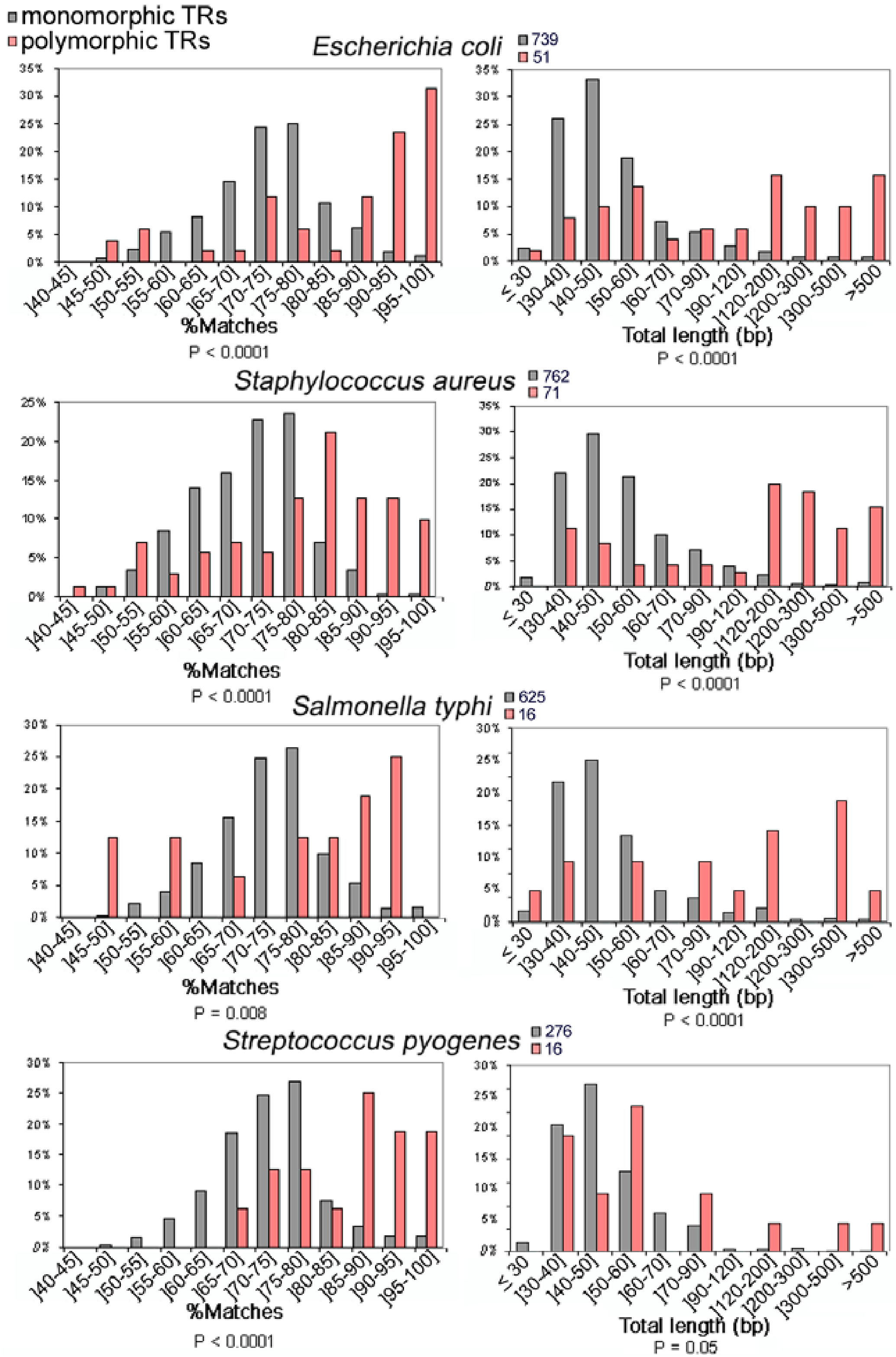



**Figure 6**
Proportion of predicted polymorphic (pink) and monomorphic (grey) tandem repeats according to different parameters (internal homogeneity of the repeat array (%matches) or total length). P-values obtained for the non-parametric Wilcoxon tests appear below each histogram.

**Table 1: Use of the criterion "total length ≥ 80 bp and %matches ≥ 80%" on 4 species for which 3 strains or more were compared. The number of monomorphic, polymorphic (2 alleles or more) and highly polymorphic (3 alleles or more) TRs in whole set, and positive and negative groups are listed. (a) "criterion" refers to the selection of TRs with L ≥ 80 bp and %M ≥ 80%**

| Comparison (total number of TRs) | Whole set (proportion of total number) | | | Tandem repeats with L≥80 bp AND %M≥80% (proportion among the set) | | | Tandem repeats with L<80 bp OR %M<80% (proportion among the set) | | | % of the polymorphic TRs (2 alleles or more) that were detected by criterion[a] | % of the TRs with 3 alleles or more that were detected by criterion[a] | % of all TRs that fulfil the criterion[a] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 allele | 2 alleles or more | 3 alleles or more | 1 allele | 2 alleles | 3 alleles or more | 1 allele | 2 alleles or more | 3 alleles or more | | | |
| *S aureus* (833 TRs) | 762 (91.5%) | 71 (8.5%) | 38 (4.5%) | 5 (13%) | 8 (20%) | 25 (66%) | 757 (95%) | 25 (3.5%) | 13 (1.5%) | 46% | 66% | 7.23% |
| *E coli* (790 TRs) | 739 (93.5%) | 51 (6.5%) | 12 (1.5 %) | 12 (38%) | 13 (40%) | 7 (22%) | 727 (96%) | 26 (3.5%) | 5 (0.5%) | 39% | 58% | 4.86% |
| *S typhi / typhimurium* (641 TRs) | 625 (97.5%) | 16 (2.5%) | 2 (0.3%) | 13 (68%) | 4 (22%) | 2 (10%) | 612 (98%) | 10 (2%) | 0 (0%) | 37.5% | 100% | 3.27% |
| *S pyogenes* (292 TRs) | 276 (94.5%) | 16 (5.5%) | 3 (1%) | 4 (67%) | 0 (0%) | 2 (33%) | 272 (95%) | 14 (4.7%) | 1 (0.3%) | 12.5% | 67% | 2.71% |

**Table 2: Use of the criterion "total length ≥ 80 bp and %matches ≥ 80%" on 10 species for which 2 strains were compared. The numbers of tandem repeats with equal lengths and different lengths between the two strains in the whole set, and positive and negative groups are listed.**

| Comparison (total number of TRs loci) | Whole set (proportion) | | Criterion + (L≥80 bp, %M≥80%) | | Criterion - | | Sensitivity (% of the TRs with different lengths that were detected by criterion) | Specificity (% of the TRs predicted by the criterion that have different length) | % of all TRs that fulfil the criterion |
|---|---|---|---|---|---|---|---|---|---|
| | equal length | different length | equal length | different length | equal length | different length | | | |
| *H pylori* 26695/J99 (624 TRs) | 506 (81%) | 118 (19%) | 0 | 11 | 506 | 107 | 9% | 100% | 2% |
| *N meningitidis* MC58/Z2491 (642 TRs) | 528 (82%) | 114 (18%) | 10 | 23 | 518 | 91 | 20% | 70% | 5% |
| *M tuberculosis* H37Rv/ CDC1551(1502 TRs) | 1441 (96%) | 61 (4%) | 35 | 27 | 1406 | 34 | 44% | 44% | 4% |
| *L monocytogenes* EGD-e/*L innocua* Clip11262 (576 TRs) | 553 (96%) | 23 (4%) | 2 | 3 | 551 | 20 | 13% | 60% | 1% |
| *S agalactiae NEM316/2603* (398 TRs) | 387 (97%) | 11 (3%) | 2 | 1 | 385 | 10 | 9% | 33% | 1% |
| *S pneumoniae* TIGR4/R6 (406 TRs) | 339 (83%) | 67 (17%) | 14 | 29 | 325 | 38 | 43% | 67% | 10% |
| *Y pestis CO-92/KIM5 P12* (1499 TRs) | 1372 (92%) | 127 (8%) | 44 | 19 | 1328 | 108 | 15% | 30% | 4% |
| *R prowazekii* Madrid E/*R conorii* malish 7 (316 TRs) | 290 (92%) | 26 (8%) | 0 | 2 | 290 | 24 | 8% | 100% | 1% |
| *Brucella suis* 1330/ *Brucella melitensis* 16 M (739 TRs) | 681 (92%) | 58 (8%) | 2 | 4 | 679 | 54 | 7% | 67% | 1% |
| *X fastidiosa* 9a5c/grape Temecula1 (573 TRs) | 440 (77%) | 133 (23%) | 2 | 28 | 438 | 105 | 21% | 93% | 5% |

predictive criteria. For all four species, the group fulfilling the criterion is, as expected, enriched in polymorphic (at least two alleles) tandem repeats: in *Staphyloccocus aureus*, polymorphic tandem repeats represent only 8.5% of the whole population of tandem repeat loci but are predominant (87%) in the criterion positive group. The enrichment is even greater for highly polymorphic TRs, *i.e.* with 3 alleles or more: for example from 4.5% in the whole set to 66% in the positive group for *Staphylococcus aureus*. However this simple criterion misses more than half of the polymorphic loci. In addition, the efficiency of the criterion is highly variable in the different species: it is relatively satisfying in *Staphylococcus aureus* (54% of polymorphic tandem repeats would be missed) but very inefficient in *Streptococcus pyogenes* (almost 90% are missed). The results for highly polymorphic loci (3 alleles or more) are more consistent (the proportion of TRs with 3 alleles or more detected by the criterion ranges from 58% for *Escherichia coli* to 100% for *Salmonella*).

It is tempting to speculate that these observations are applicable to other species. Subsequently, we applied the criterion to ten of the 2-strains comparisons available on the Strain Comparison Page (Table 2). In all ten instances, the criterion positive group is enriched in TRs with different lengths between the two strains, compared to the whole set. This proportion varies from less than 3% in *Streptococcus agalactiae* to more than 20% in *Xylella fastidiosa* in the whole set. It is increased to 33% and 93% respectively among the set of loci which satisfy the criterion (these percentages correspond to the predictor's specificity), but the vast majority of polymorphic loci will be missed (90% and 80% respectively). Sensitivity, that is % of the TRs with different lengths that were detected by criterion varies from 6.90% for *Brucella* to 44.26% for *Mycobacterium tuberculosis*.

The finding that polymorphic tandem repeats have, on average, a higher internal conservation, total length, and copy number than monomorphic ones is in agreement with previous observations that TR polymorphism is correlated with conservation in *Yersinia pestis* and with total length in *Bacillus anthracis* [9]. It is also reminiscent of the behavior of microsatellites (also called short sequence repeats: SSR, see [45] for review), which are stabilized by internal variations [46] and by reduction of the number of repeats [47]. Unfortunately, we show here that such simple prediction criteria may miss a very large proportion of polymorphic tandem repeats, and provide highly variable results in different species. This indicates that, in the absence of sequence data from two strains or more, the systematic testing of tandem repeats polymorphism across a set of relevant strains remains the most appropriate way to develop an MLVA assay. Consequently, the Strain Comparison page is of great use when two strains or more have been sequenced.

## Conclusions

Bacterial strain typing at the subspecies level is essential for epidemiological issues in the context of disease control. This can be used to determine if an *S. aureus* or *P. aeruginosa* infection for instance has been acquired in an hospital environment or not. On a larger scale, it can be used to trace the emergence of new, more virulent or drug resistant *M. tuberculosis* strains. It is also of interest in the field of bioterrorism and bioweapons control, as was shown by the investigations following the 2001 *B. anthracis* attacks. Tandem repeats typing has recently emerged as one way to address this issue. Indeed, in the case of a number of highly monomorphic bacterial species, including *B. anthracis* and *Y. pestis*, tandem repeats typing is the method of choice for subspecies typing. In addition to the fact that these loci represent an important fraction of the existing polymorphism, it offers a number of practical advantages, including the ease of typing, and of data exchanges among different countries. It is hoped that the tools which are described here will help evaluate the potential of tandem repeats typing assays for a larger range of pathogens.

## Availability

All the tools presented are freely available from http://minisatellites.u-psud.fr.

## List of abbreviations used

ASP: active server pages

MLVA: multiple locus VNTR analysis

PCR: polymerase chain reaction

TR: tandem repeat

TRF: tandem repeats finder

## Authors contributions

FD is the developer of the database and web site, and the curator of the database. GV participated in the development of the initial procedure for the tandem repeat size comparisons between two genomes. The two authors contributed equally to the writing.

## Acknowledgments

This work was funded by grants from Délégation Générale de l'Armement (DGA, France) aimed at facilitating the typing of dangerous pathogens.

## References

1. van Belkum A: **High-throughput epidemiologic typing in clinical microbiology.** *Clin Microbiol Infect* 2003, **9:**86-100.
2. Enright MC, Spratt BG: **Multilocus sequence typing.** *Trends Microbiol* 1999, **7:**482-7.
3. Gutacker MM, Smoot JC, Migliaccio CA, Ricklefs SM, Hua S, Cousins DV, Graviss EA, Shashkina E, Kreiswirth BN, Musser JM: **Genome-wide analysis of synonymous single nucleotide polymorphisms in *Mycobacterium tuberculosis* complex organisms: resolution of genetic relationships among closely related microbial strains.** *Genetics* 2002, **162:**1533-43.
4. Achtman M, Zurth K, Morelli G, Torrea G, Guiyoule A, Carniel E: ***Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*.** *Proc Natl Acad Sci U S A* 1999, **96:**14043-8.
5. van Belkum A, Scherer S, van Leeuwen W, Willemse D, van Alphen L, Verbrugh H: **Variable number of tandem repeats in clinical strains of *Haemophilus influenzae*.** *Infect Immun* 1997, **65:**5017-27.
6. Frothingham R, Meeker-O'Connell WA: **Genetic diversity in the *Mycobacterium tuberculosis* complex based on variable numbers of tandem DNA repeats.** *Microbiology* 1998, **144:**1189-1196.
7. Supply P, Mazars E, Lesjean S, Vincent V, Gicquel B, Locht C: **Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome.** *Mol Microbiol* 2000, **36:**762-771.
8. Adair DM, Worsham PL, Hill KK, Klevytska AM, Jackson PJ, Friedlander AM, Keim P: **Diversity in a variable-number tandem repeat from *Yersinia pestis*.** *J Clin Microbiol* 2000, **38:**1516-9.
9. Le Flèche P, Hauck Y, Onteniente L, Prieur A, Denoeud F, Ramisse V, Sylvestre P, Benson G, Ramisse F, Vergnaud G: **A tandem repeats database for bacterial genomes: application to the genotyping of *Yersinia pestis* and *Bacillus anthracis*.** *BMC Microbiol* 2001, **1:**2.
10. Le Flèche P, Fabre M, Denoeud F, Koeck JL, Vergnaud G: **High resolution, on-line identification of strains from the *Mycobacterium tuberculosis* complex based on tandem repeat typing.** *BMC Microbiol* 2002, **2:**37.
11. Spanier JG, Jones SJ, Cleary P: **Small DNA deletions creating avirulence in *Streptococcus pyogenes*.** *Science* 1984, **225:**935-8.
12. Hollingshead SK, Fischetti VA, Scott JR: **Size variation in group A streptococcal M protein is generated by homologous recom-**

**bination between intragenic repeats.** *Mol Gen Genet* 1987, **207:**196-203.
13. Keim P, Kalif A, Schupp J, Hill K, Travis SE, Richmond K, Adair DM, Hugh-Jones M, Kuske CR, Jackson P: **Molecular evolution and diversity in *Bacillus anthracis* as detected by amplified fragment length polymorphism markers.** *J Bacteriol* 1997, **179:**818-24.
14. Keim P, Price LB, Klevytska AM, Smith KL, Schupp JM, Okinaka R, Jackson PJ, Hugh-Jones ME: **Multiple-Locus Variable-Number Tandem Repeat Analysis Reveals Genetic Relationships within *Bacillus anthracis*.** *J Bacteriol* 2000, **182:**2928-2936.
15. Sylvestre P, Couture-Tosi E, Mock M: **Polymorphism in the collagen-like region of the *Bacillus anthracis* BclA protein leads to variation in exosporium filament length.** *J Bacteriol* 2003, **185:**1555-63.
16. Supply P, Lesjean S, Savine E, Kremer K, van Soolingen D, Locht C: **Automated high-throughput genotyping for study of global epidemiology of *Mycobacterium tuberculosis* based on mycobacterial interspersed repetitive units.** *J Clin Microbiol* 2001, **39:**3563-3571.
17. Vergnaud G, Denoeud F: **Minisatellites: Mutability and Genome Architecture.** *Genome Res* 2000, **10:**899-907.
18. Pourcel C, Vidgop Y, Ramisse F, Vergnaud G, Tram C: **Characterization of a Tandem Repeat Polymorphism in *Legionella pneumophila* and Its Use for Genotyping.** *J Clin Microbiol* 2003, **41:**1819-1826.
19. Onteniente L, Brisse S, Tassios PT, Vergnaud G: **Evaluation of the polymorphisms associated with tandem repeats for *Pseudomonas aeruginosa* strain typing.** *J Clin Microbiol* 2003, **41:**4991-7.
20. **GOLD Genomes OnLine Database** [http://ergo.integratedgenomics.com/GOLD/]
21. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27:**573-580.
22. **The Tandem Repeats Finder** [http://tandem.bu.edu/trf/trf.html]
23. Denoeud F, Vergnaud G, Benson G: **Predicting Human Minisatellite Polymorphism.** *Genome Res* 2003, **13:**856-867.
24. **The ActiveState Programmer Network (ASPN) ActivePerl download page** [http://www.activestate.com/products/ActivePerl]
25. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-402.
26. **The NCBI BLAST ftp site** [ftp://ftp.ncbi.nih.gov/blast/]
27. **The tandem repeats database** [http://minisatellites.u-psud.fr]
28. **The Strain Comparison Page** [http://minisatellites.u-psud.fr/comparison]
29. Parkhill J, Wren BW, Thomson NR, Titball RW, Holden MT, Prentice MB, Sebaihia M, James KD, Churcher C, Mungall KL, Baker S, Basham D, Bentley SD, Brooks K, Cerdeno-Tarraga AM, Chillingworth T, Cronin A, Davies RM, Davis P, Dougan G, Feltwell T, Hamlin N, Holroyd S, Jagels K, Karlyshev AV, Leather S, Moule S, Oyston PC, Quail M, Rutherford K, Simmonds M, Skelton J, Stevens K, Whitehead S, Barrell BG: **Genome sequence of *Yersinia pestis*, the causative agent of plague.** *Nature* 2001, **413:**523-7.
30. Deng W, Burland V, Plunkett G 3rd, Boutin A, Mayhew GF, Liss P, Perna NT, Rose DJ, Mau B, Zhou S, Schwartz DC, Fetherston JD, Lindler LE, Brubaker RR, Plano GV, Straley SC, McDonough KA, Nilles ML, Matson JS, Blattner FR, Perry RD: **Genome sequence of *Yersinia pestis* KIM.** *J Bacteriol* 2002, **184:**4601-11.
31. Radnedge L, Agron PG, Worsham PL, Andersen GL: **Genome plasticity in *Yersinia pestis*.** *Microbiology* 2002, **148:**1687-98.
32. Fleischmann RD, Alland D, Eisen JA, Carpenter L, White O, Peterson J, DeBoy R, Dodson R, Gwinn M, Haft D, Hickey E, Kolonay JF, Nelson WC, Umayam LA, Ermolaeva M, Salzberg SL, Delcher A, Utterback T, Weidman J, Khouri H, Gill J, Mikula A, Bishai W, Jacobs WR Jr, Venter JC, Fraser CM: **Whole-Genome Comparison of *Mycobacterium tuberculosis* Clinical and Laboratory Strains.** *J Bacteriol* 2002, **184:**5479-5490.
33. Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, Han CG, Ohtsubo E, Nakayama K, Murata T, Tanaka M, Tobe T, Iida T, Takami H, Honda T, Sasakawa C, Ogasawara N, Yasunaga T, Kuhara S, Shiba T, Hattori M, Shinagawa H: **Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12.** *DNA Res* 2001, **8:**11-22.
34. Perna NT, Plunkett G 3rd, Burland V, Mau B, Glasner JD, Rose DJ, Mayhew GF, Evans PS, Gregor J, Kirkpatrick HA, Posfai G, Hackett J, Klink S, Boutin A, Shao Y, Miller L, Grotbeck EJ, Davis NW, Lim A, Dimalanta ET, Potamousis KD, Apodaca J, Anantharaman TS, Lin J, Yen G, Schwartz DC, Welch RA, Blattner FR: **Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7.** *Nature* 2001, **409:**529-33.
35. Welch RA, Burland V, Plunkett G 3rd, Redford P, Roesch P, Rasko D, Buckles EL, Liou SR, Boutin A, Hackett J, Stroud D, Mayhew GF, Rose DJ, Zhou S, Schwartz DC, Perna NT, Mobley HL, Donnenberg MS, Blattner FR: **Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*.** *Proc Natl Acad Sci U S A* 2002, **99:**17020-4.
36. **The Blast in the tandem repeats database page** [http://minisatellites.u-psud.fr/Blast]
37. **The Bacterial Genotyping Page** [http://bacterial-genotyping.igmors.u-psud.fr]
38. Tettelin H, Masignani V, Cieslewicz MJ, Eisen JA, Peterson S, Wessels MR, Paulsen IT, Nelson KE, Margarit I, Read TD, Madoff LC, Wolf AM, Beanan MJ, Brinkac LM, Daugherty SC, DeBoy RT, Durkin AS, Kolonay JF, Madupu R, Lewis MR, Radune D, Fedorova NB, Scanlan D, Khouri H, Mulligan S, Carty HA, Cline RT, Van Aken SE, Gill J, Scarselli M, Mora M, Iacobini ET, Brettoni C, Galli G, Mariani M, Vegni F, Maione D, Rinaudo D, Rappuoli R, Telford JL, Kasper DL, Grandi G, Fraser CM: **Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V *Streptococcus agalactiae*.** *Proc Natl Acad Sci U S A* 2002, **99:**12391-6.
39. McClelland M, Sanderson KE, Spieth J, Clifton SW, Latreille P, Courtney L, Porwollik S, Ali J, Dante M, Du F, Hou S, Layman D, Leonard S, Nguyen C, Scott K, Holmes A, Grewal N, Mulvaney E, Ryan E, Sun H, Florea L, Miller W, Stoneking T, Nhan M, Waterston R, Wilson RK: **Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2.** *Nature* 2001, **413:**852-6.
40. Jin Q, Yuan Z, Xu J, Wang Y, Shen Y, Lu W, Wang J, Liu H, Yang J, Yang F, Zhang X, Zhang J, Yang G, Wu H, Qu D, Dong J, Sun L, Xue Y, Zhao A, Gao Y, Zhu J, Kan B, Ding K, Chen S, Cheng H, Yao Z, He B, Chen R, Ma D, Qiang B, Wen Y, Hou Y, Yu J: **Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157.** *Nucleic Acids Res* 2002, **30:**4432-41.
41. Bricker BJ, Ewalt DR, Halling SM: **Brucella 'Hoof-Prints': strain typing by multi-locus analysis of variable number tandem repeats (VNTRs).** *BMC Microbiol* 2003, **3:**15.
42. Goodner B, Hinkle G, Gattung S, Miller N, Blanchard M, Qurollo B, Goldman BS, Cao Y, Askenazi M, Halling C, Mullin L, Houmiel K, Gordon J, Vaudin M, Iartchouk O, Epp A, Liu F, Wollam C, Allinger M, Doughty D, Scott C, Lappas C, Markelz B, Flanagan C, Crowell C, Gurson J, Lomo C, Sear C, Strub G, Cielo C, Slater S: **Genome sequence of the plant pathogen and biotechnology agent *Agrobacterium tumefaciens* C58.** *Science* 2001, **294:**2323-8.
43. Wood DW, Setubal JC, Kaul R, Monks DE, Kitajima JP, Okura VK, Zhou Y, Chen L, Wood GE, Almeida NF Jr, Woo L, Chen Y, Paulsen IT, Eisen JA, Karp PD, Bovee D Sr, Chapman P, Clendenning J, Deatherage G, Gillet W, Grant C, Kutyavin T, Levy R, Li MJ, McClelland E, Palmieri A, Raymond C, Rouse G, Saenphimmachak C, Wu Z, Romero P, Gordon D, Zhang S, Yoo H, Tao Y, Biddle P, Jung M, Krespan W, Perry M, Gordon-Kamm B, Liao L, Kim S, Hendrick C, Zhao ZY, Dolan M, Chumley F, Tingey SV, Tomb JF, Gordon MP, Olson MV, Nester EW: **The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58.** *Science* 2001, **294:**2317-23.
44. Benson G, Dong L: **Reconstructing the duplication history of a tandem repeat.** *Proc Int Conf Intell Syst Mol Biol* 1999:44-53.
45. van Belkum A, Scherer S, van Alphen L, Verbrugh H: **Short-sequence DNA repeats in prokaryotic genomes.** *Microbiol Mol Biol Rev* 1998, **62:**275-93.
46. Schumacher S, Fuchs RP, Bichara M: **Two distinct models account for short and long deletions within sequence repeats in *Escherichia coli*.** *J Bacteriol* 1997, **179:**6512-7.
47. De Bolle X, Bayliss CD, Field D, van de Ven T, Saunders NJ, Hood DW, Moxon ER: **The length of a tetranucleotide repeat tract in *Haemophilus influenzae* determines the phase variation rate of a gene with homology to type III DNA methyltransferases.** *Mol Microbiol* 2000, **35:**211-22.